Inteligencia Artificial, 21(62) (2018), 91-102
doi: 10.4114/intartif.vol21iss62pp91-102

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Deep Learning Applied on Refined Opinion Review Datasets

Ingo Jost[1], João Francisco Valiati[2]
[1]CWI Software, São Leopoldo, Brazil
ingo.jost@gmail.com
[2]Artificial Intelligence Engineers-AIE, Porto Alegre, Brazil
joao.valiati@ai-engineers.com

Abstract Deep Learning has been successfully applied in challenging areas, such as image recognition and audio classification. However, Deep Learning has not yet reached the same performance when employed in textual data classification, including Opinion Mining. In models that implement a deep architecture, Deep Learning is characterized by the automatic feature selection step. The impact of previous data refinement in the pre-processing step before the application of Deep Learning is investigated to identify opinion polarity. The refinement includes the use of a classical procedure of textual content and a popular feature selection technique. The results of the experiments overcome the results of the current literature with the Deep Belief Network application in opinion classification. In addition to overcoming the results, their presentation is broader than the related works, considering the change of parameter variables. We prove that combining feature selection with a basic preprocessing step, aiming to increase data quality, might achieve promising results with Deep Belief Network implementation.

Keywords: Deep Learning; Opinion Mining; Feature Selection; Deep Belief Networks.

1 Introduction

The continuous growth of data volume contributes to the improvement of techniques that seek the implicit knowledge of these data. The Knowledge Discovery in Database (KDD) area is furthered by the technological advances in recent years, standing out in its performance in different approaches such as Text Mining, a specific field of KDD that treats pattern recognition in textual data, like document classification[3].

When these textual data are about opinions, specific points arise and they are treated by Opinion Mining. It is a specialization of Text Mining that helps the decision making process and allows companies to understand what customers think about their products[4], and customers take better choices based on the purchase experience from other buyers. Different models have been applied in opinion classification problems. Among them stands out Deep Learning, which is employed in several fields of pattern recognition like image[10] and audio[14] identification, character classification[15], and face recognition[13]. The significant results in these areas open possibilities to apply Deep Learning in other fields like text mining. However, the Deep Learning application in opinion classification does not overcome the current literature results[1], motivating this investigation.

A major quality of Deep Learning is the feature selection from data[31], where the algorithm learns multiple levels of data representation and the learned representations can be considered as features[47].

ISSN: 1137-3601 (print), 1988-3064 (on-line)

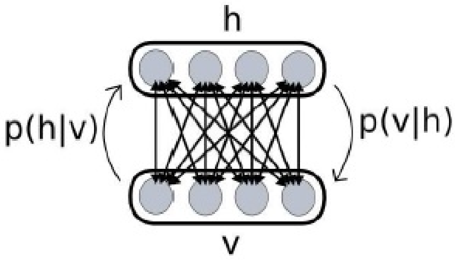

Figure 1: RBM's structure[20]

The aim of this research is to provide an analysis of the impact of data refinement, the use of a classical text pre-processing and a feature selection technique, exerts on polarity classification with the Deep Belief Network (DBN)[24], a specialization of Deep Learning. Although most researches use Deep Learning in raw data[16], this study demonstrates the benefits of pre-processing techniques.

The main contibution of this works is to demonstrate the effectiveness of the combination between classical preprocessing, with a feature selection method, and DBN can provide competitive results when compared with the current literature (see Table 6). Deep Learning, with focus on DBN, and Opinion Mining are introduced in the next sections, followed by related works that build models with deep architectures to apply on review data. After, the experiments and the obtained results are related. Finally, the conclusion is presented.

2 Deep Learning

Machine Learning is a research field open to the development and extension of methods with a continuous search for best practices, considering costs and results. This search favours the development of new areas, such as Deep Learning. Its concepts are based on Artificial Neural Networks (ANN)[49], having biological inspiration in the human brain and studies about the mammalian visual cortex[9].

In recent years researchers have tried to increase the quantity of layers of ANNs[27]. Success was not achieved until 2006, when an algorithm was proposed to train Deep Belief Networks[24]. This algorithm uses multiple layers composed of non-linear information for feature selection (supervised or unsupervised), transformation and pattern analysis, trying to identify relationships in the data[31].

The use of unsupervised learning algorithms to learn features from unlabeled data has contributed to Deep Learning[21] growth, evidencing its skill in feature selection and distributed representation on multiple levels[32]. Through various researches, the classifiers have been lead to achieve better results when using Deep Learning.

These results were also obtained in various Text Mining approaches: semantic and sense identification from terms[22], text clustering[3] and domain adaptation of Opinion Mining[23]. Ain et al.[52] presented a review of distinct approaches of Deep Learning in Opinion Mining. Similarly, Zhang et al.[53] provided a survey of Deep Learning applications on tasks of Sentiment Analysis. They were implemented by different architectures such as Stacked Auto-encoder, Convolutional Neural Networks[30, 51] and Deep Belief Networks.

The DBNs are probabilistic models composed by one visible and many hidden layers[30]. Each of these is compounded by Restricted Boltzmann Machines (RBM), learning statistical relationships among the lower level layers. The RBMs are a specific case of Boltzmann Machines formed by visible (v) and hidden (h) units, with the restriction of forbidden connections between neurons of the same layer[21], according to Figure 1. Their structure is bipartite graph, consisting of a stochastic neural network with only one hidden layer each, which tries to find a likeness from input data in an unsupervised way. The Boltzmann Machines (including RBMs) are Energy Based Models (EBM), associating an energy scalar for each joint configuration (pairs of visible and hidden units)[27]. This energy value allows calculating the distribution p from vector v, used for unit computing and updating of weights[20].

The DBN layers are formed by stacked RBMs hierarchically arranged, which are individually trained[19],

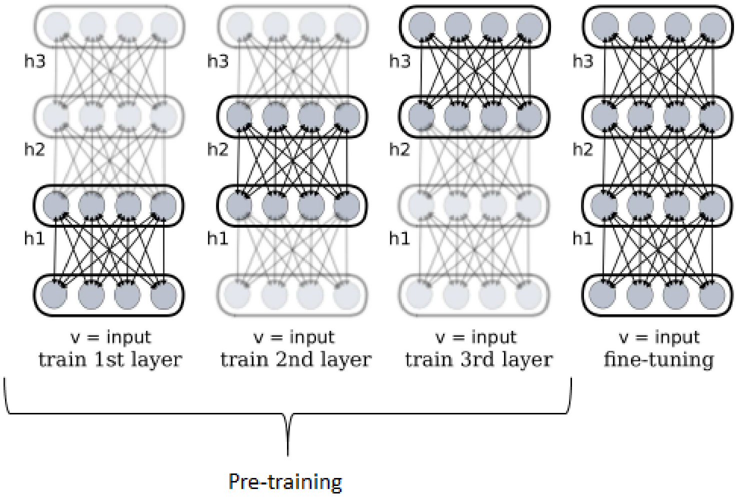

Figure 2: The DBN's training steps[20]

and the hidden units from each RBM are the visible units for the next layer (except for the last layer). This step, in which the RBMs are trained layer-by-layer, is called Pre-training. Thereafter, the whole model has the weights adjusted by backpropagation in the Fine-tuning stage. These steps characterize the algorithm proposed by Hinton et al.[24] and are illustrated in Figure 2, where each RBM is highlighted during its training in the Pre-training and the layers are connected in Fine-tuning.

The number of epochs in the Pre-training and Fine-tuning steps and the number of neurons for hidden layers in deep architectures are empirically determined, such as in shallow architectures. However, the researches that normally apply DBN use three hidden layers, with no references that recommend a higher number. Our work followed the premises of most of the works and employed three hidden layers (see Experimental Setup section).

3 Opinion Mining

Opinion Mining is a Text Mining specialization that deals with opinions[26]. Its growth is favoured by the web 2.0 scenario and social media content[25], a collaborative environment that makes possible the development of tools that allow users to send their opinion through discussion groups, forums, social networks, blogs, websites news, product sales, etc. Furthermore, the corporations increasingly need to understand the feelings of customers about their products and services[11]. This information helps in the decision making process and strengthens the need for studies on Opinion Mining.

Several challenges appear when the terms can have different meanings and a positive review can have terms frequently used in negative sentences, such as subjectivity. Moreover, there are sentiments that are very hard to identify, like irony and sarcasm, and the contradiction of ideas with the use of negation. In addition to these questions, the handling of the occurrence of implicit and conditional opinions is needed[26]. Although opinions belong to different domains (products or news) and are generated by different tools, studies for classifying opinions are often directed toward a common aim: polarity identification (opinion classification). This consists of identifying the class to which an opinion refers: positive or negative[11].

Since the Opinion Mining treats textual data, some specific procedures for handling texts are performed [43]: tokenization, the identifying words (tokens) in a text[44]; stemming, to group different terms but with the same radical[45]; and stopwords removal. Stopwords are terms that often occur in text but do not get to contribute to the process of classification or pattern recognition. In order to reduce the volume, these words can be removed without harming the meaning of the text. Besides, the feature selection (term selection) procedure is applied to reduce the data volume and enhance their quality. Among the techniques for feature selection, such as Probability Ratio and Chi-squared, the IG is a popular feature selection approach[5], with a low computational cost and it has reached competitive results in the feature selection[41, 42]. The IG makes use of entropy and calculating the information gain for each term. More information about data is provided through this metric, enabling improvement in their quality and turning them into a refined form. These procedures are part of the pre-processing step.

After pre-processing there is a transformation step, in which the opinions are transformed into word vectors, where each position represents a term and storage value that will be used by the classifier algorithm. This value is a weight representation, such as TF-IDF[34] or the term frequency. In this way, the data is suited to the classifiers, like DBN, to be trained. The next section presents related works that implemented these applications using Deep Learning techniques.

4 Related Works

According to Xia et al.[46], the studies involving Opinion Mining usually follow a traditional text classification, where the documents are mapped into a feature vector through the bag-of-words (BOW) model and posteriorly classified by Machine Learning techniques, like Naïve Bayes (NB) or Support Vector Machines (SVM). Many works apply neural networks[1] with shallow architectures. Although there are recent studies that investigate several flavors of Deep Learning applied to Sentiment Analysis like exposed in [53], our focus is on the works that investigated the same datasets, such as the works presented in this section.

The work presented in Zhou et al.[7] uses the Active Learning model to classify opinions, comparing their results with the experiments done by other authors with the Deep Belief Networks implemented by Hinton[24]. Glorot et al.[9] develops a Deep Learning model with Denoising Autoencoder[27], adopting a rectifier function (and its smooth form, the Softplus function[12]) for neuron activation to image recognition and applies the same model to opinion classification. Both works produce experiments with the same datasets investigated in our work (movies and books reviews) and their results are explored as a benchmark. Glorot et al.[9] is an exception because their results are reported as an average of four datasets.

The following sections present the specifications of these works and determined aspects not appropriately explored, like the networks parameterization.

4.1 Active Learning

Zhou et al.[7] proposed the Active Deep Network, a semi-supervised learning method to classify opinions through Deep Learning architecture. The experiments were performed for datasets from Pang and Lee[8] and on another four datasets of different types of products from Amazon: books (BOO), DVDs (DVD), electronics (ELE), and kitchen's appliances (KIT). Each set was formed by 2,000 opinions equally divided into positive and negative classes.

The pre-processing was similar to Dasgupta[28], where vectors of unigrams represent opinions. The terms were selected by frequency, discarding the 1.5% used more frequently (assuming that these represent stopwords). The training was called semi-supervised because only part of the samples were labeled to adjust the weights. In the Pre-training step, all samples participated in the layer-by-layer unsupervised training architecture to generate the weight matrix. The supposed most difficult samples to classify were chosen from the training set through the technique of Active Learning, in order to consider the label for supervised training. This difficulty level was measured by the distance of the sample from the class separator on the hyperplane (the closer to the separator, the more difficult it was to classify). Finally, the model training considered the selected samples.

The authors performed the experiments with 10-folds randomly divided, which were tested with cross-validation. The results were compared with the other referenced models, highlighting the reached accuracy by Active Deep Learning (ADN) for books and movies datasets, 69% and 76%, respectively, overcoming the experiments of the authors with the original DBN for the same datasets (64% and 71%).

An extension for ADN was proposed in Zhou et al.[2]: the Information ADN (IADN), which used the information's density to choose the samples that will pass for supervised learning. Instead of considering just the distance of the point from the separator point on the hyperplane, it also considered the distance from the center of the classes.

The experiments reached similar results to those obtained by the ADN when 100 labeled samples were used for the five datasets. However, the IADN produced better results when 10 labeled samples were used. Just one configuration of neurons per layer was used, the last layer being the output and the previous three layers corresponding to the hidden layers: 100-100-200-2. Furthermore, the training occurred with only 30 epochs.

The model demonstrated in this section improved the DBN implementation. The results were presented with just one setup of neurons per hidden layer, one number of epochs, and a fixed number of terms. In our work, an original DBN implementation was suited and the obtained accuracy was higher than experiments from Zhou et al.[7]. Moreover, it analysed the parameters change and its influence on results.

4.2 Rectifier Neural Network

Glorot et al.[9] proposed the use of the rectifier activation function for neurons, analysing the effects of using the Pre-training step in their Deep Learning framework. The work was an extension of Nair and Hinton[10].

The rectifier function generated a sparse representation of the data, according to biological inspiration, since studies have shown that neurons store information sparsely[29]. An advantage of this representation was the possibility of keeping data variability. When very dense, small changes can affect their vector representation[27]. Furthermore, sparse data have a tendency to be linearly separable.

Four datasets were used: MNIST, images of digits; CIFAR10, RGB images; NISTP, character images; and Norb, picture of toys. The experiments were realized with and without the Pre-training step, comparing the results. The experiments that performed the Pre-training step achieved the best results.

In addition to the sets of images, the model was applied to classify sentiments in reviews from the OpenTable website (10,000 labeled opinions and 300,000 not labeled). Each review was reduced through the BOW model and converted to binary vector, representing the presence or absence of the term. The 5,000 most frequent terms were considered. The experiments using the Pre-training step produced higher accuracy than experiments without Pre-training, proving its effectiveness.

Since no publication was related to the OpenTable data, the authors applied the rectifier neural network following the pre-processing setup defined by Zhou et al.[7], with the same datasets, using the four datasets from Amazon (BOO, DVD, ELE and KIT). The work reached an average accuracy of 78.9%, overcoming the 73.7% obtained by Zhou et al.[7]. Additionally, the results achieved by Ghosh et al.[54], which combined a two layered RBM to dimensionality reduction and a Probabilistic Neural Network (PNN), and Ruangkanokmas et al.[48], whom applied the Deep Belief Network to the Feature Selection (DBNFT) model, combining the Chi-square feature selection technique with common pre-processing before the DBN application, are also demonstrated. Table 1 presents the average accuracy obtained in the discussed works.

Table 1: Accuracy average (%) in related works in BOO, DVD, ELE and KIT datasets

Model	Average
DBN (Zhou et al. [7])	69.3
ADN (Zhou et al. [7])	73.7
IADN (Zhou et al. [2])	64.5
DBNFT (Ruangkanokmas et al. [48])	71.4
RNN (Glorot et al.[9])	78.9
PNN (Ghosh et al.[54])	80.1

These works demonstrated that Deep Learning offers different possibilities for application in opinion classification. Recent works have applied Deep Learning to obtain promising results in Sentiment Analysis considering data from social networks, such as the following propositions: Deep Recurrent Neural

Networks [36], Deep Memory Networks [38] and the proposed of Wang et al. [37], which combined Convolutional Neural Network and Recurrent Neural Network (CNN+RNN). Despite these last approaches, our work showed that the application of more effort in the pre-processing step lead to greater accuracy. The obtained accuracy of related works was presented for further comparison with our results (see Comparison of Results section).

5 Experiments

This investigation used two datasets: the classical dataset of opinions about movies from Pang and Lee[8] and opinions about books from Amazon. The employed evaluation made use of the 10-fold cross-validation method, and over each set of training and tests generated by these folds preprocessing and classification tasks were applied. The IG method was applied to find more representative terms after the use of a basic clean procedure: tokenization, removal stopwords list, and stemming (Snowball algorithm). We followed the premises of Moraes et al.[5], which created sets with different amounts of terms, then we evaluated the respective number of terms (300, 500, 1,000, 2,000, 3,000, 4,000, 5,000). Different from the related works[2, 7, 9], we also analyse the extension of network parameters, like the number of neurons by layer.

The Pre-training stage was perceived in the studies related with textual data as an unsupervised feature selector[3, 23], following the approaches of other fields, like image recognition[50]. When these applications were made for opinion classification, a basic pre-processing, considering the frequency of terms, was applied[9, 7]. The purpose of applying IG for term selection in our investigation was to identify the influence that a basic and popular dimensionality reduction technique exerts over DBN. Before effectively applying the DBN, the data were transformed into vectors containing the Term Frequency from each selected term.

The implementation of a classifier with a deep architecture by Ruslan Salakhutdinov and Geoffrey Hinton[1] was used for experimenting. This classifier was a framework that implemented Deep Learning concepts, which trained a DBN with three hidden layers formed by RBMs. These layers were individually trained in the Pre-training stage and then the weights were adjusted in the Fine-tuning step.

The original framework's code was adapted to train and test opinion data, including modifications in input and output layers due to domain specification. An architecture with three hidden layers was maintained for all experiments, according to related works. The steps are illustrated in Figure 3. After the pre-processing (which prepared the refined data) and training steps, the model produced an output for each sample of the test set. This output was compared with the intended values, generating the confusion matrix, formed by the TP, TN, FP, and FN values.

The True Positives (TP) were the positive samples correctly classified and True Negatives (TN) the negative samples predicted as negative. The False Positives (FP) and False Negatives (FN) corresponded to the sample amount wrongly predicted as positive and negative, respectively[43]. From these metrics is calculated the accuracy (Eq. (1)) used in the related works and in our results.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

5.1 Experimental Setup

Maintaining a similar architecture found in related works, three hidden layers were adopted. The initial settings used in Zhou et al.[7] were replicated, i.e., three hidden layers with 100, 100, and 200 neurons and 30 epochs for training. Preliminary experiments demonstrated that a greater number of neurons per layer and more epochs contributed to improve the accuracy.

Seeking a suitable configuration of nodes in hidden layers that could produce better accuracy, experiments were realized beginning with the configuration from Zhou et al.[7], and gradually and proportionally the number of neurons was increased. These settings of neurons per hidden layer were used: 100, 100, 200; 200, 200, 400; 300, 300, 600; 400, 400, 800; and 500, 500, 1,000.

[1] http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html

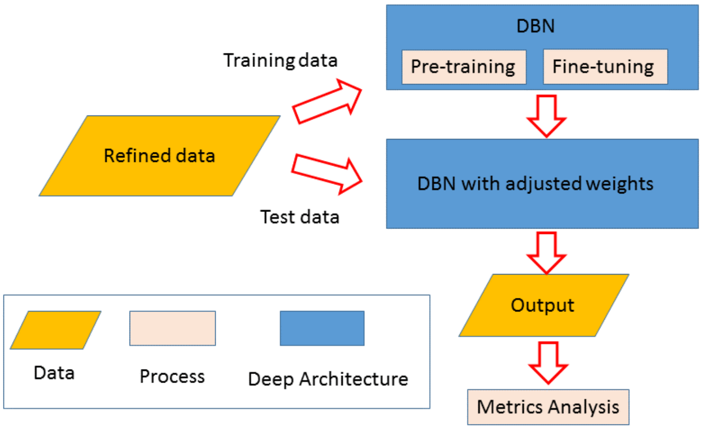

Figure 3: The flowchart of experiments

Since the Pre-training stage followed an unsupervised step, the end of training procedures was determined by the number of epochs. After preliminary tests, we chose to realize experiments with a higher number of epochs for training if compared to Zhou et al.[7]. We followed the approaches that applied Deep Learning in textual data, in which experiments were done with fewer epochs in the Pre-training[3], preventing a high number of epochs in this unsupervised step harmed the training. Different amounts of epochs were analysed for the Pre-training (PT) and Fine-tuning (FT) steps: 120 (PT) and 120 (FT); 160 (PT) and 160 (FT); and 30 (PT) and 120 (FT).

5.2 Obtained Results

Experiments were realized in movies and books datasets, varying the configuration of the three hidden layers and the sets with a different number of terms, as previously specified. Applying the configuration with 120 epochs in the Pre-training and Fine-tuning steps, the maximum accuracy obtained was 82% for movies (considering 2,000 terms) and 77.7% for books (considering 500 terms).

We repeated the same configurations with 160 epochs in Pre-training and Fine-tuning steps to check whether or not the continuous increase of epochs contributed for reaching better results. The accuracy obtained is presented in Tables 2 and 3, where it is verified that the movies dataset accuracy was greater in most of the experiments (comparing to experiments with 120 epochs in the Pre-training and Fine-tuning), while for books the results were closer to previous experiments.

Table 2: Accuracy (%) for movies - 160 epochs in Pre-training and Fine-tuning steps

Layers	Number of terms						
1st, 2nd, 3rd	300	500	1,000	2,000	3,000	4,000	5,000
100, 100, 200	79.0	75.9	75.2	75.5	75.2	75.2	75.3
200, 200, 400	**81.6**	81.9	82.0	80.8	80.4	77.3	76.3
300, 300, 600	81.1	**82.2**	**82.4**	**82.8**	**82.4**	82.3	**82.6**
400, 400, 800	80.7	81.8	82.1	82.7	**82.4**	82.1	82.5
500, 500, 1,000	79.6	81.1	82.1	82.3	**82.4**	**82.7**	82.2

Table 3: Accuracy (%) for books - 160 epochs in Pre-training and Fine-tuning steps

Layers	Number of terms						
1st, 2nd, 3rd	300	500	1,000	2,000	3,000	4,000	5,000
100, 100, 200	77.0	77.4	**76.8**	75.8	75.9	74.9	75.3
200, 200, 400	**77.3**	**77.7**	76.7	76.3	76.3	75.6	75.6
300, 300, 600	77.1	76.7	76.5	76.5	75.8	76.4	75.8
400, 400, 800	76.0	76.7	76.4	**76.6**	**76.6**	75.9	**76.0**
500, 500, 1,000	76.2	76.6	76.2	76.5	76.2	**76.5**	75.4

Although higher accuracy was obtained for the movies dataset, the improvement was not significant. For this reason, we adopted the strategy from Salakhutdinov and Hinton[3] that reduced the number of epochs in the Pre-training step. This was done in order to avoid unsupervised training with a high number of epochs, which caused the overfitting problem[39].

The obtained results in experiments with 30 epochs in Pre-training and 120 in Fine-tuning are shown in Tables 4 (movies) and 5 (books). The accuracy for the movies dataset did not exceed the obtained results in previous experiments (with 120 and 160 epochs in both steps). However, the strategy of a smaller number of epochs reached greater accuracy for the books dataset.

Table 4: Accuracy (%) for movies - 30 epochs in Pre-training and 120 in Fine-tuning

Layers	Number of terms						
1st, 2nd, 3rd	300	500	1,000	2,000	3,000	4,000	5,000
100, 100, 200	79.1	75.1	75.0	74.3	74.2	73.8	73.9
200, 200, 400	**81.1**	**81.6**	**81.5**	80.6	79.1	77.2	76.5
300, 300, 600	80.6	81.0	81.3	81.3	81.3	**81.5**	81.4
400, 400, 800	80.0	80.6	81.4	81.4	**81.5**	**81.5**	**81.7**
500, 500, 1,000	78.3	80.0	80.8	**81.8**	81.1	81.0	81.5

Table 5: Accuracy (%) for books - 30 epochs in Pre-training and 120 in Fine-tuning

Layers	Number of terms						
1st, 2nd, 3rd	300	500	1,000	2,000	3,000	4,000	5,000
100, 100, 200	**77.3**	77.6	**77.2**	76.3	75.8	75.1	**76.0**
200, 200, 400	77.0	**77.8**	**77.2**	76.4	**76.6**	75.8	75.5
300, 300, 600	76.6	76.7	76.7	**77.2**	76.4	**76.5**	75.9
400, 400, 800	75.9	76.4	76.3	76.1	75.6	76.2	75.7
500, 500, 1,000	76.0	75.9	76.1	75.5	75.7	75.7	74.9

In spite of the lower accuracy achieved for the movies dataset, the obtained results were competitive with the previous configurations. Moreover, when the Pre-training step with a smaller number of epochs was realized, the computational demand was significantly lower, turning this configuration to the recommended setting.

The data refinement strategy produced satisfactory results for the most experiments and settings, overcoming the related works, as discussed in the next section.

5.3 Comparison of Results

In Zhou et al.[7], the proposed Active Deep Learning (ADN) model applied the frequency for the terms selection and reached 76.3% accuracy for the movies dataset and 69% for the books dataset, overcoming the 71.3% and 64.3% (movies and books, respectively) obtained with the same experiments with DBN implementation. The ADN results were adopted as a benchmark in the IADN[2], Rectifier Neural Network (RNN)[9] and Ruangkanokmas et al.[48].

In this investigation, a complementary experiment was realized submitting the movies reviews in raw format to the AlchemyAPI, an online commercial tool, recently aquired by IBM Watson, that receives documents and identifies through Deep Learning implementation (without informing the applied algorithm) whether or not they indicate a positive or negative sentiment.

Table 6 shows that the presented work achieved the best accuracy when compared with the results produced by Zhou et al.[7] in experiments with original DBN, with the ADN model, and its improvement (IADN)[2]. Moreover, the obtained accuracy in the AlchemyAPI experiments, the results from DBNFT[48], PNN[54] and CNN+RNN[37] were also related. The results of Glorot et al.[9] were not included because their presentation was an average (78.9%) of all datasets, considering different data from our work.

Table 6: Comparison of Results - Accuracy(%)

Model	Movies	Books
DBN Zhou et al. [7]	71.3	64.3
ADN Zhou et al. [7]	76.3	69.0
IADN Zhou et al. [2]	76.4	69.7
AlchemyAPI	77.8	-
DBNFT Ruangkanokmas et al. [48]	72.2	66.0
PNN Ghosh et al. [54]	80.8	81.0
CNN+RNN Wang et al. [37]	82.3	-
Present work	82.8	77.8

Although our DBN implementation did not exceed the results obtained in opinion classification researches, such as Bai[17] beating the 90% accuracy for the movies dataset from Pang and Lee[8], the strategy of refining data in the pre-processing step before the DBN application overcame the results found in related works that had applied a simple pre-processing. Moreover, the obtained accuracy was higher than the accuracy produced by the experiments with the AlchemyAPI using the raw data. Besides to exceeding the results, some network parameters were analysed to verify their impact in the deep architecture. Comparing Tables 2 and 4 for movies and Tables 3 and 5 for books, the variation of these parameters did not produce statistically significant improvements.

6 Conclusion

It was observed that the obtained results for opinion classification with the use of Deep Belief Networks, like reported by the current literature, do not exceed the results obtained with classical data mining techniques. However, this investigation opens possibilities of different approaches of Deep Learning application, such as the previous refinement of the data and the analysis of different parameters for the classifiers. The use of refinements in the original datasets, like the application of pre-processing techniques and feature selection[5], combined with the Deep Belief Network implementation for training and classifying of polarity classes, helps reach promising results.

Although the works compared from the current literature used a basic pre-processing, they only considered the frequency of terms. Our investigation proved that the strategy with data refinement - applying IG - allowed achieving and even overcoming the obtained results, increasing the accuracy in 6% and 8% for movies and books datasets, respectively. The experiments were realized with a wide setup, including a variety of parameterization. These experiments produced close results, confirming the successful data refinement strategy for the most experiments. Future extensions of the proposed work conducted with the use of this methodology to other datasets related to opinion classification, and the possibility to investigate a particular extension of Deep Learning, called Recursive Neural Tensor Network (RNTN)[6], have recently been recommended for Sentiment Analysis.

7 Acknowledgments

We thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

References

[1] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems* (2015). http://dx.doi.org/10.1016/j.knosys.2015.06.015

[2] S. Zhou, Q. Chen, X. Wang, "Active deep learning method for semi-supervised sentiment classification," *Neurocomputing* **120** (2016) 536–546.

[3] R. Salakhutdinov and G. Hinton, "Semantic hashing," International Journal of Approximate Reasoning **50** (2010) 969–978. doi:10.1016/j.ijar.2008.11.006

[4] S. Basari, B. Hussin, G. P. Ananta and J. Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization," *Procedia Engineering* **53** (2013) 453–462. doi:10.1016/j.proeng.2013.02.059

[5] R. Moraes, J. F. Valiati and W. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Appli cations* **40** (2013) 621–633. doi:10.1016/j.eswa.2012.07.059

[6] R. Socher, A. Perelygin, J. Y. Wu, C. D. Manning, A. Ng and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1631–1642.

[7] S. Zhou, Q. Chen and X. Wang, "Active deep networks for semi-supervised sentiment classification," *23rd International Conference on Computational Linguistics, Association for Computational Linguistics*, 2010, pp. 1515–1523.

[8] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," *42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics,* 2004, pp. 271–278. doi: 10.3115/1218955.1218990

[9] X. Glorot, A. Bordes and Y. Bengio, "Deep Sparse Rectifier Neural Networks," *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[10] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," *27th International Conference on Machine Learning*, 2010, pp. 807–814.

[11] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.

[12] C. Dugas, Y. Bengio, F. Belisle, C. Nadeau and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," *Advances in Neural Information Processing Systems*, 2001, pp. 472–478.

[13] H. Fan, Z. Cao, Y. Jiang, Q. Yin and C. Doudou, *Learning Deep Face Representation*, Cornell University Library, Computer Vision and Pattern Recognition, 2014. arXiv preprint arXiv:1403.2802

[14] M. Norouzi, *Convolutional Restricted Boltzmann Machines for Feature Learning*, Simon Fraser University, 2009.

[15] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* **313** (2006) 504–507.

[16] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems* **19** (2007) 153–160.

[17] X. Bai, “Predicting consumer sentiments from online text, ” *Decision Support Systems* **4** (2011) 732–742. doi:10.1016/j.dss.2010.08.024

[18] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.

[19] G. Hinton, “Deep Belief Nets,” *NIPS Tutorial, Canadian Institute for Advanced Research and Department of Computer Science University of Toronto*, 2007.

[20] L. Arnold, S. Rebecchi, S. Chevallier and H. P. Moisy, “An introduction to deep-learning,” *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2011.

[21] Y. Yang, “Learning Hierarchical Representations for Video Analysis Using Deep Learning,” Ph. D. Thesis, University of Central Florida, 2013.

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Ng and C. Potts, “Learning Word Vectors for Sentiment Analysis,” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 1*, 2011, pp. 142–150.

[23] X. Glorot, A. Bordes and Y. Bengio, *Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach*, Omnipress, 2011.

[24] G. Hinton, S. Osindero and Y. M. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation* **18** (2006) 1527–1554.

[25] H. Chen and D. Zimbra, “AI and Opinion Mining,” *IEEE Intelligent Systems* **25** (2010) 74–80. doi:10.1109/MIS.2010.75

[26] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval* **2** (2008) 1–135.

[27] Y. Bengio, *Learning Deep Architectures for AI*, Now Publishers Inc, Boston, 2009.

[28] S. Dasgupta and V. Ng, “Mine the Easy, Classify the Hard, A Semi-supervised Approach to Automatic Sentiment Classification,” *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009.

[29] D. Attwell and S. B. Laughlin, “An Energy Budget for Signaling in the Grey Matter of the Brain,” *Journal of Cerebral Blood Flow and Metabolism* **21** (2001) doi:10.1097/00004647-200110000-00001

[30] H. Lee, P. T. Pham, Y. Largman and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[31] L. Deng and D. Yu, *Deep Learning: Methods and Applications*, NOW Publishers, 2014.

[32] R. Socher, Y. Bengio and C. D. Manning, “Deep learning for NLP (without magic),” *Tutorial Abstracts of ACL 2012, Association for Computational Linguistics*, 2012.

[33] Y. He, C. Lin and H. Alani, “Automatically Extracting Polarity-bearing Topics for Cross-domain Sentiment Classification,” *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1, 2011, pp. 123–131.

[34] G. Paltoglou and M. Thelwall, “A Study of Information Retrieval Weighting Schemes for Sentiment Analysis,” *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1386–1395.

[35] A. Yessenalina, Y. Yue and C. Cardie, “Multi-level Structured Models for Document-level Sentiment Classification,” *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 1046–1056.

[36] C. Li, X. Guo and Q. Mei, "Deep Memory Networks for Attitude Identification" *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 671-680. doi: 10.1145/3018661.3018714

[37] X. Wang, W. Jiang, Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts" *Proceedings of the International Conference on Computational Linguistics*, 2016, doi: 10.1145/3155133.3155158

[38] Z. Zhao, H. Lu, D. Cai, X. He and Y. Zhuang, "Microblog Sentiment Classification via Recurrent Random Walk Network Learning" *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3532–3538. doi: 10.24963/ijcai.2017/494

[39] C. Schittenkopf, G. Deco and W. Brauer, "Two Strategies to Avoid Overfitting in Feedforward Networks," *Neural Networks* **10** (1997) 505–516. doi:10.1016/S0893-6080(96)00086-X

[40] N. V. Chawla, "Data Mining for Imbalanced Datasets, An Overview," *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Springer, 2010, pp. 853–867.

[41] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Fourteenth International Conference on Machine Learning*, 1997, pp. 412–420.

[42] A. Abbasi, S. France, Z. Zhang and H. Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," *IEEE Transactions on Knowledge and Data Engineering* **23** (2011) 447–462. doi:10.1109/TKDE.2010.110

[43] S. M. Weiss, N. Indurkhya and T. Zhang, *Text mining. Predictive methods for analyzing unstructured information*, Springer, 2004.

[44] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2006.

[45] M. W. Sholom, I. Nitin and Z. Tong, *Fundamentals of Predictive Text Mining*, Springer, 2006.

[46] S. M. Weiss, N. Indurkhya and T. Zhang, "Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification," *Information Sciences* **181** (2011) 1138–1152. doi:10.1016/j.ins.2010.11.023

[47] D. Tang, B. Qin and T. Liu, "Deep learning for sentiment analysis: successful approaches and future challenges," *WIREs Data Mining Knowl Discov* **5** (2015) 292–303. doi: 10.1002/widm.1171

[48] P. Ruangkanokmas, T. Achalakul, K. Akkarajitsakul, "Deep Belief Networks with Feature Selection for Sentiment Classification," *7th International Conference on Intelligent Systems, Modelling and Simulation*, 2016.

[49] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.

[50] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," 2015. arXiv 1512.03385v1

[51] A. Mishra, K. Dey, P. Bhattacharyya, "Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network,", Proceedings of the Annual Meeting of the Association for Computational Linguistics 2017. doi: 10.18653/v1/P17-1035

[52] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat and A. Rehman, "Sentiment Analysis Using Deep Learning Techniques: A Review", *International Journal of Advanced Computer Science and Applications* **6** 2017. doi: 10.14569/IJACSA.2017.080657

[53] L. Zhang, S. Wang, B. Liu, "Deep Learning for Sentiment Analysis: A Survey", *WIREs Data Mining Knowl Discov.* **8** 2018. doi: 10.1002/widm.1253

[54] R. Ghosh, K. Ravi, V. Ravi, "A novel deep learning architecture for sentiment classification," *3rd Int'l Conf. on Recent Advances in Information Technology* 2016. doi: 10.1109/RAIT.2016.7507953